JOINT ATTENTION IN HUMAN-ROBOT INTERACTION

A Thesis
Presented to
The Academic Faculty

by

Chien-Ming Huang

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
College of Computing

Georgia Institute of Technology
August 2010

JOINT ATTENTION IN HUMAN-ROBOT INTERACTION

Approved by:

Professor Andrea L. Thomaz,
Committee Chair
College of Computing
Georgia Institute of Technology

Professor Rosa I. Arriaga
College of Computing
Georgia Institute of Technology

Professor Henrik I. Christensen
College of Computing
Georgia Institute of Technology

Date Approved: 6 July 2010

To my parents and my brother,

for their unconditional support, encouragement, and love.

ACKNOWLEDGEMENTS

I want to thank my advisor Dr. Andrea Thomaz, without whose inspiration and encouragement I would never have started this journey. She is one of the most supportive mentors I have encountered. I also want to thank Dr. Rosa Arriaga and Dr. Henrik Christensen for offering critical advice; their insightful feedback and suggestions helped shape the development and completion of this thesis. I also want to thank my labmates in the Socially Interactive Machine lab for their help and feedback on this thesis. Last but not the least, I want to thank my family members for their support and encouragement to me during the process of completing my degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Joint attention, a crucial component in interaction and an important milestone in human development, has drawn a lot of attention from the robotics community recently. Robotics researchers have studied and implemented joint attention for robots for the purposes of achieving natural human-robot interaction and facilitating social learning. Most previous work on the realization of joint attention in the robotics community has focused only on responding to joint attention and/or initiating joint attention. Responding to joint attention is the ability to follow another's direction of gaze and gestures in order to share common experience. Initiating joint attention is the ability to manipulate another's attention to a focus of interest in order to share experience. A third important component of joint attention is ensuring, where by the initiator ensures that the responders has changed their attention. However, to the best of our knowledge, there is no work explicitly addressing the ability for a robot to ensure that joint attention is reached by interacting agents. We refer to this ability as ensuring joint attention and recognize its importance in human-robot interaction.

We propose a computational model of joint attention consisting of three parts: responding to joint attention, initiating joint attention, and ensuring joint attention. This modular decomposition is supported by psychological findings and matches the developmental timeline of humans. Infants start with the skill of following a caregiver's gaze, and then they exhibit imperative and declarative pointing gestures to get a caregiver's attention. Importantly, as they aged and social skills matured, initiating actions often come with an ensuring behavior that is to look back and forth between the caregiver and the referred object to see if the caregiver is paying attention to the referential object.

We conducted two experiments to investigate joint attention in human-robot interaction. The first experiment explored effects of responding to joint attention. We hypothesize that humans will find that robots responding to joint attention are more transparent, more competent, and more socially interactive. Transparency helps people understand a robot's intention, facilitating a better human-robot interaction, and positive perception of a robot improves the human-robot relationship. Our hypotheses were supported by quantitative data, results from questionnaire, and behavioral observations. The second experiment studied the importance of ensuring joint attention. The results confirmed our hypotheses that robots that ensure joint attention yield better performance in interactive human-robot tasks and that ensuring joint attention behaviors are perceived as natural behaviors by humans. The findings suggest that social robots should use ensuring joint attention behaviors.

CHAPTER I

INTRODUCTION

People have envisioned that robots, after the inventions of personal computers and the Internet, are the next revolutionary technology that will change people's daily lives in the foreseeable future. With more and more research and industrial efforts put into robotics technology, robots of various functionalities and purposes (toy robots: Aibo, Pleo; service robots: Roomba; and therapeutic robots: Keepon, Paro) are available to the public. The emergence of robotics technology in the daily lives of people has brought several interesting and challenging questions to the robotics community. What kind of modalities and mechanisms do robots need to live in a human environment? How should robots behave and interact with people in ways that meet people's expectations and adhere to social norms? How can people without any training or knowledge about robotics interact with robots in a similar way as they interact with other people? To answer these questions, robotics researchers study psychology, cognitive science, and neuroscience to learn how people develop social skills and why people behave in the ways they do. One fundamental key to these questions is the capacity for social attention coordination [4], also known as joint attention that is a process to share one's current attention with another by using social cues such as gaze.

Joint attention is a crucial component in interactions and an important milestone in human development. A common characteristic of people on the autism spectrum is difficulties in communication and interaction with other people, and it is hypothesized that the failure to develop normal joint attention abilities is at the root of this deficiency [3, 7]. Therefore, to facilitate natural human-robot interaction, we believe

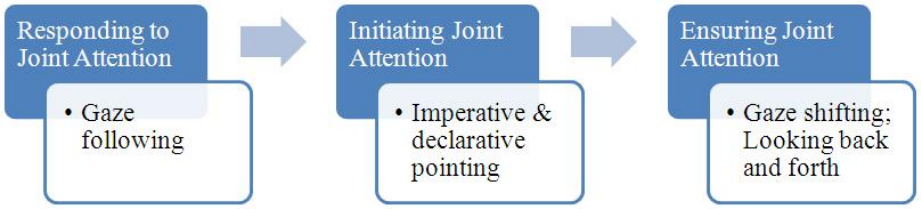

Figure 1: Developmental flow of joint attention.

that robots need social skills to respond to, initiate, and maintain joint attention with humans.

We propose a model of joint attention reflecting the complexity of this social skill for cognitive robots, extending the account in [24] by dividing the skill into its three main parts: responding to joint attention (RJA), initiating joint attention (IJA), and ensuring joint attention (EJA).

- RJA is the ability to follow another's direction of gaze and gestures in order to attain common experience.
- IJA is the ability to manipulate another's attention to a focus of interest in order to share experience.
- EJA is the ability to ensure that joint attention is reached.

These three parts conceptually match with the developmental milestones of joint attention in infancy (Figure 1). Infants start with the skill of following a caregiver's gaze at the sixth month, and then by the ninth and the twelfth month they exhibit imperative and declarative pointing gesture respectively to get the caregiver's attention. It is important to note that as they aged and their social skills matured the initiating actions often come with an ensuring behavior that is to look back and forth between the caregiver and the referred object. The purpose of this gaze shifting is to make sure the caregiver is attending to the right focus. The design of the model is also supported by findings in [24, 39].

There have been many works on how to realize joint attention on embodied platforms. Most previous works followed two frameworks: (1) building a constructive or learning model of developmental joint attention [12, 13, 15, 25] and (2) building a computational model of joint attention [16, 19, 22, 29, 34]. The first framework is based on building a learning model to acquire the skill of joint attention through interactions with human partners. The second framework adopts a modular-based approach to pre-program a computational model of joint attention. This thesis work pursues the second framework. However, most works adopting the second framework focus on aspects of RJA and IJA only. To the best of our knowledge, our work is the first to explicitly address the importance of ensuring joint attention in facilitating human-robot interaction. In addition, we highlight the relationship between RJA, IJA, and EJA.

The contribution of this work is threefold. First, we examine joint attention in human-robot interaction in a more comprehensive way than prior work. Instead of only focusing on RJA and/or IJA, we view joint attention as an integration of RJA, IJA, and EJA. Second, we investigate the effects of responding to joint attention on human-robot interaction in a teaching scenario. Our hypotheses that subjects have a better mental model of a robot, have a better understanding of a robot's current state, and perceive a robot more competent and socially interactive if the robot exhibits RJA were supported by data from experimental interactions. The findings suggested that robots responding to joint attention are more transparent to humans. In addition, we believe that the positive perception of robots will improve human-robot relationship. Third, we recognize and establish the importance of ensuring joint attention in human-robot interaction. Experimental results showed that EJA behaviors yield better performance in a human-robot interactive task. Moreover, people perceive EJA behaviors as natural behaviors that humans do, and they would like robots to have EJA behaviors for facilitating human-robot interaction.

The structure of this thesis is organized as follows. In chapter 2, we review attention and joint attention from a psychology perspective. We then present how our joint-attention model fits with developmental psychology findings. In chapter 3, we review related work on realizing joint attention in interactive systems. We then present the concepts and implementation of our joint-attention model in chapter 4. An experiment of investigating effects of responding to joint attention on human-robot interaction is reported in chapter 5. In chapter 6, we look into the importance of ensuring joint attention in human-robot interaction. Finally, we conclude with a discussion of our findings from the experiments and future directions for joint attention and social robots.

CHAPTER II

JOINT ATTENTION

In this chapter, I first compare the differences between attention and joint attention and examine the development of joint attention in human infancy. I then present research findings showing the potential existence of joint attention in non-human primate. Finally, I present some research works and issues regarding to joint attention.

2.1 Attention and Joint Attention

Attention is a process to focus on some features of the environment while ignoring others. According to [18], attention can be categorized as either passive attention or active attention. Passive attention occurs when a salient event, such as a sharp sound, happens. Active attention occurs when an agent is involved in an intentionally directed process, and the agent needs to selectively focus on certain features in the environment to achieve a particular situation. For example, driving requires active attention from people to selectively focus on traffic. Moreover, attention can be described as directed perception (e.g., eye gaze) [38]. For instance, people pay attention to something by looking at or leaning toward (in the case of hearing) it. Furthermore, attention was regarded as goal-driven directed perception in [18, 35].

From a physiological point of view, orienting reflex is concerned with attention. Orienting reflex is a fundamental change of behavior involving turning the eyes, head and body toward an alarming external stimulus [26]. The concept and the biological advantage of orienting reflex first described by Pavlov was noted in [31]:

> "It is the reflex which brings about the immediate response in man and animals to the slightest changes in the world around them, so that they

> immediately orientate their appropriate receptor-organ in accordance with the perceptible quality in the agent bringing about the change, making full investigation of it. The biological significance of this reflex is obvious. If the animal were not provided with such a reflex its life would hang at every moment by a thread. In man this reflex has been greatly developed in its highest form by inquisitiveness-the parent of that scientific method through which we hope one day to come to a true orientation in knowledge of the world around us."

Joint attention, however, is defined as a process to share one's current interest in the environment with others by using social cues, such as gaze or pointing gestures. According to [7], joint attention involves a triadic relationship among agent, self, and an object. In contrast, to have or to maintain attention is a process concerned with self and the environment (aspect of interest). Further explained in [19], joint attention involves two agents not only focusing on the same object but also having mutual acknowledgement of the sharing action. Moreover, Kaplan argued that joint attention is more than gaze following and simultaneous looking, instead joint attention implies viewing other agents' behaviors as intentionally-driven [18]. These implications of joint attention shares similar concepts of Tomasello's account of the development of theory of mind to social cognition [35]: children understand other person in terms of their thoughts and beliefs.

2.2 Joint Attention and Infant Development

Joint attention, also known as shared attention, has been recognized as an important milestone in infant development. In addition, it has been widely believed that joint attention is the key to social intelligence, including the ability to communicate and interact with other people. Moreover, the development of joint attention has been related to language acquisition and imitative learning [5]. It also has been suggested

that a failure to develop joint attention properly results in social deficits such as Autism Spectrum Disorder. Autistic disorder appears in the first three years of life and is characterized by impaired social interaction and communication and restricted and repetitive behaviors [3]. Children with autism have been found to have difficulties in making joint attention [17, 33, 37]. For instance, they may look at the pointing hand instead of at the object pointed to, and they may easily fail to point at objects for the purpose of commenting and sharing an experience.

The development of joint attention starts from a very early age. Findings from developmental psychology showed that normally developed infants have a special inclined to watch human faces. By the age of three months, infants are capable of maintaining eye contact. But, not until the age of nine months can infants follow eye gaze, and around the same time, infants acquire the ability to manipulate a caregiver's attention by imperative pointing [7]. Imperative pointing is used as a request for a certain object by pointing at it. Infants do imperative pointing even when a caregiver is not paying attention to them. At the age of 12 months, infants show declarative pointing, which is used to draw a caregiver's attention to a distal object such as the sun. A month later, infants start to use referential words to draw attention from a caregiver [18]. As they grow up, their ability of following eye gaze gets more matured. When they reach their second birthdays, infants can follow eye gaze toward an object outside their view.

According to [24], there are two kinds of joint-attention behavior in infancy: responding to joint attention (RJA) and initiating joint attention (IJA). RJA is the ability to respond to another's attention (such as following gaze and gestures). IJA is the ability to use eye contact and self attention to establish joint attention that facilitates later communication and interaction. A developmental timeline of joint attention with respect to RJA and IJA is summarized in Table 1. It is important to note that a person normally gazes back and forth between a referred object and the

Table 1: Developmental timeline of joint attention in infancy with respect to RJA and IJA [7, 18].

Developmental Timeline	Responding to Joint Attention	Initiating Joint Attention
Precursor of joint attention behaviors		
3 months	Eye gaze maintenance	
4 months		Ability to break gaze
6 months	Discrimination between left and right direction of head and gaze	
Joint attention behaviors		
9 months	Eye gaze following (to the first object encountered)	Imperative pointing
12 months	Eye gaze following (to the referred object); Accuracy improved if gaze is coupled with a point gesture	Declarative pointing
13 months		Referential words
24 months	Eye gaze following toward object outside of view	

other person a few times to make sure that joint attention is reached. This behavior was also observed in infancy and especially associated with IJA behaviors [24]. To address this particularly important behavioral pattern, we refer it to ensuring joint attention (EJA). Furthermore, in [39], joint attention is noted to be concerned with (1) to detect another's attentional direction (i.e., RJA), (2) to direct the attention of another (i.e., IJA), and (3) to switch gaze between an object and a person (i.e., EJA). Additionally, one aspect of EJA (i.e., the need of confirmation of eye contact) is pointed out in [16].

2.3 Joint Attention in Non-human Primate

Joint attention is also observed in great apes. In [9], Brauer et al. noted that great apes follow humans' gaze to distant locations and around barriers. In addition, all four great apes (i.e., chimpanzee, bonobo, gorilla, and orangutan) sometimes looked back to the human experimenter and double checked where the human experimenter was looking in their experiments [9]. This finding may suggest that EJA also exists in non-human primate.

It has been widely studied and believed that great apes have the ability to follow human partners' attention. However, seldom work investigated the ability of great apes to follow gaze of conspecifics. Tomasello et al. demonstrated that chimpanzees were able to follow gaze of conspecifics during experimental trials [36]. Moreover, as noted in [24], chimpanzees show the capacity for RJA but little evidence of IJA (i.e., spontaneously share experiences with conspecifics).

Hobson et al. explored the potential effect of caregiver sensitivity on joint attention in human infants and found that responsive care is correlated with increased joint attention skills [14]. A recent study on early care for great apes revealed different perspectives about the effect of early care on joint attention. Pitman and Shumaker showed that four types of great apes engaged in joint attention with humans and

conspecifics regardless of whether they received responsive or basic care from great ape mothers or humans in their first 6 months of lives [27].

2.4 Cognitive Science Research on Joint Attention

Over the years, joint attention has drawn a lot of attention from researchers in psychology, cognition science, neuroscience, and robotics. Psychologists and cognitive scientists are interested in how joint attention emerges at an early age, how joint-attention behaviors are developmentally related to one another, and how joint attention constitutes higher-level cognitive mechanisms [23, 30]. Neuroscientists study joint attention at a neural-network level by seeing which areas in a human brain are associated with joint-attention experiences and how the identified areas are related to joint attention [32, 39]. More recently, researchers in the robotics community have become interested in how to implement joint attention on robots for the purpose of facilitating human-robot interaction or achieving service or therapeutic tasks [25, 29, 34]. More robotics research on joint attention is reviewed in chapter 3.

There is a growing consensus of using embodiment platforms to study cognitive capacities. The benefit of using an embodied platform for evaluation of a computational model of joint attention has been recognized. An embodied platform provides the capability of being physically interactive and is more likely to draw natural responses from subjects. Moreover, in contrast to empirical observations, embodiment allows experiments being repeatable, and different aspects are easily separated for evaluation [18]. Also, embodiment provides access to internal states as a behavior develops [13].

CHAPTER III

RELATED WORK

There has been a lot of effort put into the study and realization of joint attention in robotics. Mostly, two different frameworks have been pursued to implement joint attention: (1) a constructive or learning approach, where a robot builds a constructive or learning model of developmental joint attention through interactions with humans [12, 13, 15, 25];and (2) a modular approach, where a robot is preprogrammed with a modular-based model of joint attention [16, 19, 22, 29, 34]. These two frameworks reflect the nurture and the nature accounts of joint attention in psychology respectively. Moreover, most works [12, 13, 15, 19, 22, 25, 34] in realizing joint attention focused on aspects of responding to joint attention (RJA) only. Some works [16, 29] addressed aspects of initiating joint attention (IJA). But, to the best of our knowledge, this is the first work to explicitly address the importance of ensuring joint attention (EJA) in facilitating human-robot interaction. Our model adopts the second framework to build a computational model of joint attention that consists of three parts: RJA, IJA, and EJA.

In this chapter, I review related work following the two frameworks and compare those related work with the present research.

3.1 Constructive Model of Developmental Joint Attention

Nagai et al. proposed a constructive model for the development of joint attention [25]. The proposed model reflected the process that an infant acquires joint attention through interacting with a caregiver. Their result suggested that the model enables a robot to acquire joint attention without any evaluation feedback from users. Moreover, with the model a robot can reproduce the staged developmental process of joint

attention (i.e., ecological stage, geometric stage, and representational stage [11]). However, the model only captured aspects of RJA and did not cover IJA and EJA.

Deák et al. reviewed two theoretical models of joint attention: Butterworth's and Baron-Cohen's model, but argued that models incorporated with multiple discrete modules are not supported by behavior evidence [13]. Therefore, they proposed a dynamic learning mechanism to realize joint attention. The mechanism learns cognitive faculties (i.e., contingency and synchrony), facial features, and motivations (i.e., drive to fixate on faces, to maintain eye contact, and to interact with caretakers) with supervision in not constructed environments (i.e., varying lighting conditions, head poses, and facial expressions). They hypothesized that the perceptual skills, learning algorithms, and internal motivations will support the emergence of gaze following (an aspect of RJA). Moreover, they emphasized the importance of using robots as testbeds to evaluate theories of joint attention.

In accordance with accounts in [13], Carlson and Triesch argued that nativist/modularist theories of joint attention are not solid to explain behavioral observations [12]. Hence, they suggested a computational model of emergence of gaze-following based on a reinforcement learning method. The suggested model resembled the RJA and was based on that gaze-following emerges from an interaction between a set of basic mechanisms: perceptual preferences, habituation, reward driven learning, and structured environment. Their result showed that gaze-following can be learned in the context of proper interaction between the proposed mechanisms.

In [15], Hoffman and colleagues proposed a probabilistic model for gaze imitation and shared attention. Specifically, they used a Bayesian model that combined saliency models to estimate a person's gaze. An instructor- and task-specific saliency model is learned by a robotic system through interactions with a human instructor. In addition to an instructor- and task-specific saliency model, they also included a prior model of saliency (i.e., color, shape, etc. of objects) into their model. With the combination of

probabilistic models, the robotic system could locate object of mutual interest more accurately over successive trials. This work only focused on gaze imitation, which concerns aspects of RJA.

3.2 Modular Model of Joint Attention

In contrast to describing joint attention a constructive or learning model, Scassellati [29] modeled joint attention mechanisms based on Baron-Cohen's developmental model, which is a discrete-module-based model [7]. Scassellati decomposed joint attention into four main stages: maintaining eye contact, gaze following, imperative pointing, and declarative pointing. These four stages implemented the eye-direction detector, the intentionality detector, and the shared attention module in Baron-Cohen's model. The four stages covered aspects of both RJA and IJA. However, the model did not ensure that another agent attends to the shared attention. Besides, Scassellati's work focused more on technical implementations of maintenance and following of eye gaze. Same as Scassellati's work, the present research does not address the aspect of the theory-of-mind module in Baron-Cohens model.

Thomaz et al. proposed a computational mechanism of shared attention that combined with emotional empathy and an affective memory system to realize social referencing [34]. Instead of adopting deictic gaze as joint attention, the proposed mechanism reflected Baron-Cohen's model where shared attention is an explicit mental state representation of appreciating what the other person's interest is about [6]. Thomaz's model differentiated referential focus from attentional focus by maintaining three foci of interest: the robot's attention, the user's attention, and the referential focus. However, their work only addressed aspects of RJA.

Imai et al. developed a speech generation system and a joint-attention mechanism for generation of situated utterances and manipulation of human behaviors [16]. The joint-attention mechanism included an eye-contact function, which establishes

a relationship between a person and a robot, and an attention expression function, which employs gaze and pointing gestures to get attention from a person. The mechanism, regarded as robot-centered joint attention, actually carried out IJA but did not address aspects of RJA, which is regarded as human-centered joint attention. Importantly, the need to ensure the achievement of eye contact (an aspect of EJA) was pointed out even though it was not implemented in their system.

Kozima and Yano described a general model for robots to acquire communicative behavior through interaction with its environment, especially with humans [19]. The model consists of three stages: being intentional (to be goal-directed), being identical (to experience other people's behavior), and being communicative (to empathetically understand other people's behavior). Their model involves joint-attention mechanism to achieve being identical with others. The joint-attention mechanism first identifies a face and then estimates face orientation/gaze direction. The robot searches the salient object in the estimated direction and maps the person's action onto own motor configuration to reproduce the joint-attention behavior (i.e., look at the object). Similar to most work, their system only covered aspects of RJA.

In [22], Marin-Urias et al. implemented joint attention using a geometric reasoning mechanism based on mental rotation and perspective taking concepts. Mental rotation is the ability to acquire the representation of the environment from another person's point of view. Perspective taking is the ability to reason from another person's point of view to obtain a representation of that person's knowledge. With the geometric reasoning mechanism, a robot is able to know not only what object a person is looking at but also what object a person cannot see. However, their work only addressed aspects of RJA.

A similar work [10] probed effects of nonverbal communication in human-robot teamwork and suggested that implicit nonverbal communication positively impacts human-robot task performance. RJA involves nonverbal social cues, such as eye gaze,

which acts as transparent communication. Our first experiment on effects of RJA confirmed their results that transparency improves human-robot task performance. However, one main difference between their work and the first experiment of the present work is that we use additional measures to test subjects' confidence in task performance.

In a recent study on engagement, Rich et al. proposed and implemented a model for recognizing engagement in human-robot interaction [28]. Engagement was defined as a process by which participants establish, maintain, and end their perceived connection during interactions they jointly undertake. The concepts of engagement have a large overlap with joint attention in interaction. Their model was based on four connection events, namely directed gaze, mutual facial gaze, adjacency pairs, and backchannels, that were identified in a human engagement study. In particular, the event of directed gaze involves aspects of IJA and RJA. Mutual facial gaze concerns aspects of EJA, and adjacency pairs are to establish connections between interacting agents (aspects of IJA and EJA). However, their work focused on recognition instead of generation of engagement behaviors.

CHAPTER IV

REALIZATION OF JOINT ATTENTION

In this chapter, I first describe an interaction scenario, where two interacting agents have a joint attention on an object, to operationalize joint attention and better illustrate the proposed model. Then, I present a conceptual model of joint attention and a computational implementation. Finally, I described an integrated system of an embodied platform and perception inputs used to study joint attention and to evaluate the proposed model.

4.1 Joint Attention: An Interaction Scenario

Let us consider joint attention in a simple interaction as in Figure 2. The interaction may be different if there are multiple objects and multiple agents. In addition, spatial configuration of objects may also influence the interaction. However, we only consider two interacting agents and one object of mutual interest for now. The interaction can be described as five steps. To start an interaction, two agents need a way to connect to each other. The purpose of this connection is to be aware of each other and to anticipate an upcoming interaction. There are several ways to establish a connection. One main way is to have an eye contact [28]. However, it is not limited to visual connections. Dialogue, specifically an adjacency pair, is also a common way to establish a connection between two interacting agents. An adjacency pair consists of two utterances by two speakers [21]. The first initiating utterance provokes the second responding utterance. In linguistics, an adjacency pair is viewed as a turn-taking in conversation. For example, consider the following dialogue.

Alan: Hey, Bob.

Bob: What's up?

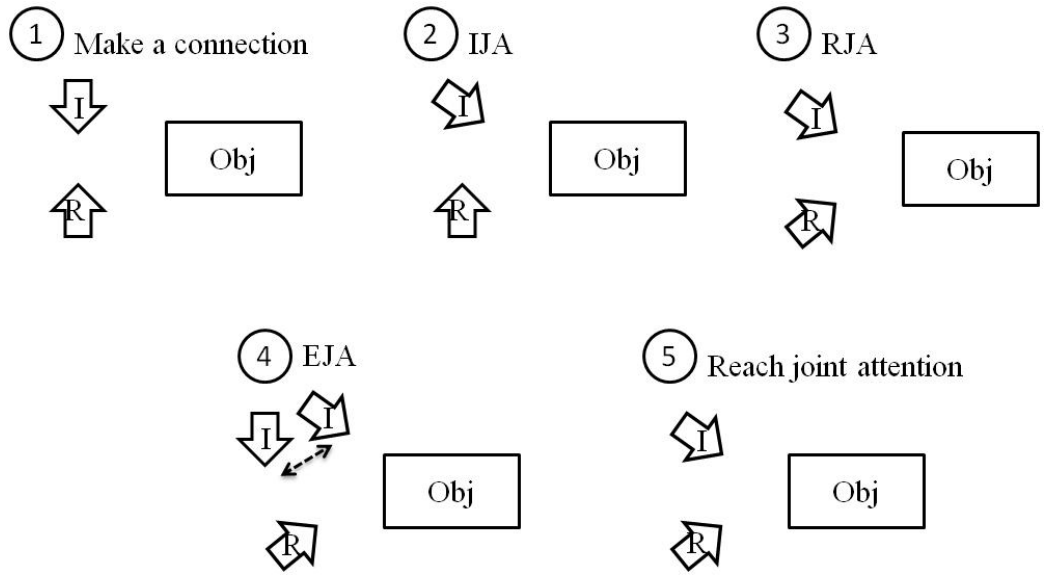

Figure 2: Joint attention in interaction. Arrow I represents an initiating agent. Arrow R represents a responding agent. Direction of an arrow indicates the agent's attentional focus.

In this conversation, Bob responds to Alan's initiating utterance and completes an adjacency pair. In this case, Alan and Bob are not necessary to see each other while speaking because they established a connection through a vocal way. However, it is normal that a visual connection (i.e., Bob walks to Alan or Alan walks to Bob) follows this kind of vocal connection. It is also common for two agents to use a combination of visual and vocal ways to establish a connection.

Once a connection is established, both agents are aware of each other and anticipate an upcoming interaction. The initiating agent makes a joint attention request by switching her attention to the object she intends to address. Switching attention (i.e., directed perception [38]) is usually to switch gaze to the focus. The initiating agent then addresses the object using communicative channels such as pointing gestures and/or vocal comments. Meanwhile, the responding agent responds to the request by switching attention to the referential focus.

Right after initiating joint attention or switching attention to the focus, the initiating agent normally looks back and forth between the responding agent and the

referential object to see if the responding agent attends to the joint attention. It is important to note that this checking process is a quick switch in gaze. Moreover, the initiating agent may do the checking and the addressing processes simultaneously (i.e., switching gaze while pointing to the focus). If the responding agent is not attending to the referential object, the initiating agent normally tries different ways to get attention from the responding agent. Ways to get attention include using bigger gestures or emphasizing gestures and making sounds.

The interacting agents reach joint attention when they both attend to the referential focus. Once joint attention is reached, the initiating agent continues the interaction (e.g., continues commenting on the referential object). At this point, both agents are focusing on the object. Importantly, the initiating agent does the ensuring joint attention process (step 4 and 5 in Figure 2) periodically during the interaction to maintain joint attention.

4.2 A Computational Model of Joint Attention

We propose a joint-attention model consisting of three parts: responding to joint attention (RJA), initiating joint attention (IJA), and ensuring joint attention (EJA) to reflect psychological findings and behavioral observations (see chapter 2). A high-level structure of the model is shown in Figure 3.

4.2.1 Responding to Joint Attention

To respond to a joint attention request, an agent gazes at or turns to the object referred to by the other agent. To do so, the agent first needs to know how the other agent conveys attention and to know where the other agent's attention is. An agent conveys attention using different methods including eye gaze, head orientation, body pose, pointing gestures, and referential words. Moreover, it is normal that an agent uses a combination of several methods at a time to draw attention from the other agent.

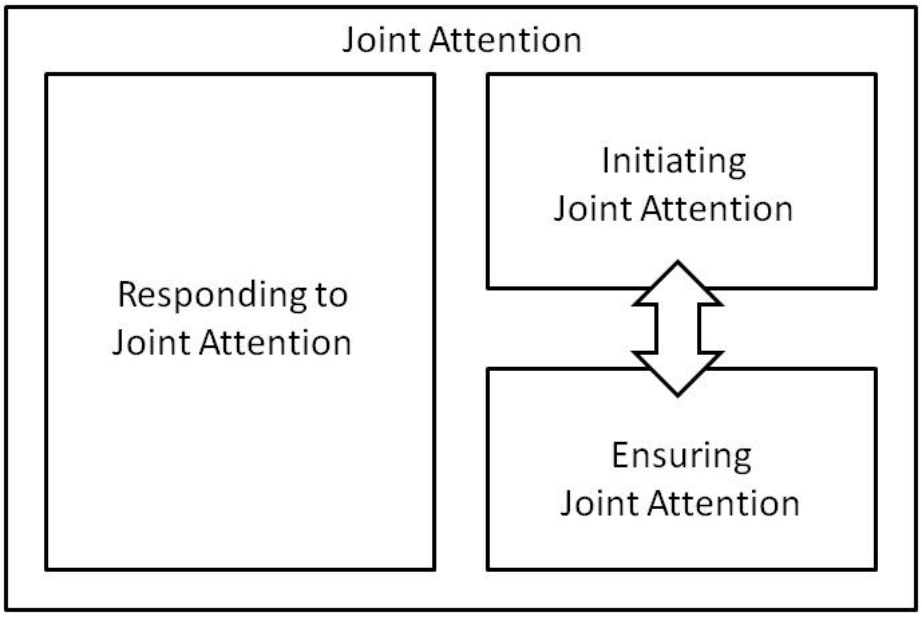

Figure 3: A high-level structure of joint-attention model consisting of three components: RJA, IJA, and EJA.

In current design, we assumed a responding agent knows the ways that an initiating agent uses to convey attention. In particular, we have implemented the RJA component to be aware of pointing gestures and referential words. Langton argued that eye gaze is not the only cue to the direction of another's attention, head orientation and pointing gestures are equivalently important [20]. Moreover, joint attention is supported by perception of multi-modal social cues. It has been shown that infants respond more to pointing gestures than gaze [13]. Nevertheless, we still would like to include gaze and head data as inputs to our model because those are important cues that humans use to convey attention. However, the technology of gaze-tracking and head-tracking in robotics is very limited in our experimental scenario. Once the data can be reliably attained it would not be hard to incorporate with our current implementation.

Table 2: An example of script of interaction

[look at the user]
R: Hey, Bob.
U: What's up.
<ja>over there </ja>(loc:0,10,50)
R: Do you see that?

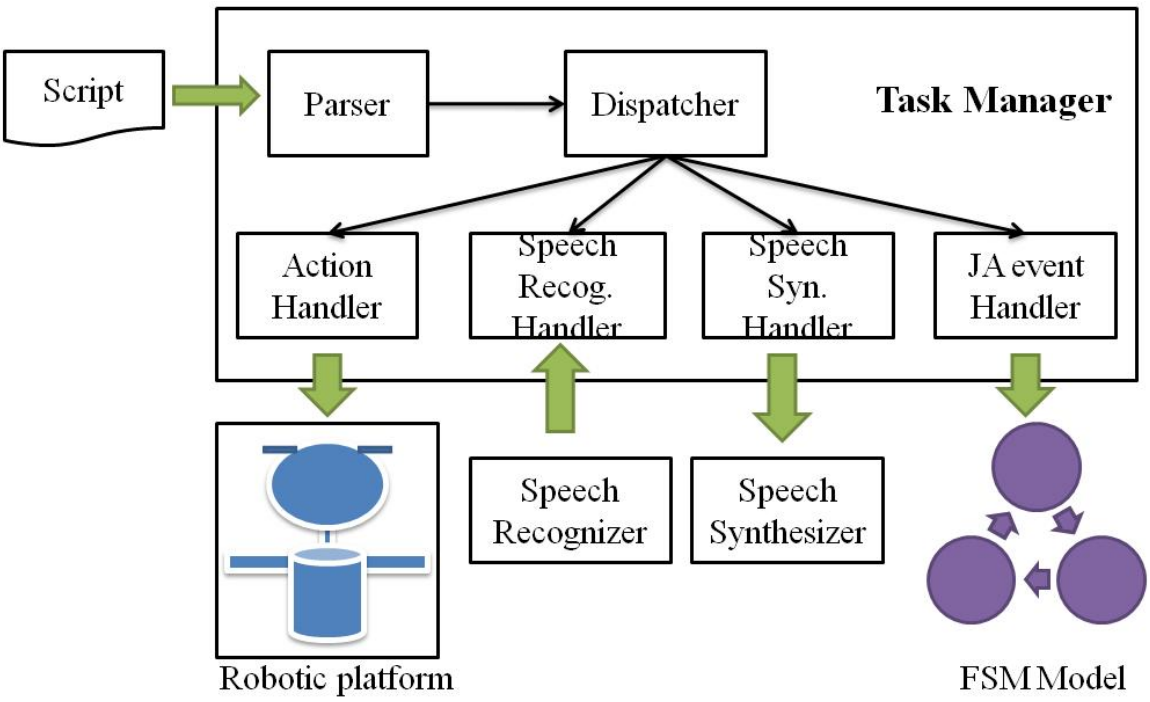

Figure 4: A system structure of script-based IJA implementation.

4.2.2 Initiating Joint Attention

An agent who intends to initiate a joint attention knows the blueprint of the interaction she is going to start. This is a reasonable assumption because if one does not know what she wants to do or show, then where and how does the will that wants to draw attention from others come from. Therefore, in our design, an initiating agent follows a script (Table 2 shows an example) that specifies actions that the agent intends to carry out ([ ]), phrases the agent wants to say (R) and the agent expects from the responding agent (U), and joint-attention event (<ja>).

In practice, an initiating agent has a task manager to carry out a script (Figure 4). The task manager first parses the script and then dispatches each command (i.e., each line) to corresponding units that handle different commands. An action command is handled by an action handler that controls and coordinates the agent's movements. Speech synthesizer and speech recognizer are responsible for self and a responding agent's phrases respectively. A joint-attention event is handled by the integrated model of IJA and EJA, which is described in section 4.2.3.

4.2.3 Ensuring Joint Attention

In our model, RJA and IJA are mutually exclusive. That is at any moment either RJA or IJA is on. However, EJA is an always-on process interacting with IJA to ensure that the other agent is attending to the right focus. Figure 5 shows the integrated model of IJA and EJA.

Recall the interaction scenario in Section 4.1. To initiate joint attention, an agent starts with establishing a connection to the other agent. The importance and the need of establishing a connection between interacting agents were pointed out in [16, 32]. A set of connecting strategies are designed to ensure a connection is established before further interaction. Connecting strategies include utterances (e.g., 'Excuse me') and using bigger gestures (e.g., waving).

Once a connection is established, the agent addresses the focus with communicative channels. The goal is to orient the other agent's attention to the addressed focus so that the two agents reach joint attention. We designed a set of communicative channels (i.e., addressing strategies) for relocating the other agent's attention. Addressing strategies include eye gaze, pointing gestures, and utterances. According to [13, 20], cues other than eye gaze are also important to predict attention of the other agent. This implies that we can also use similar cues to draw attention from other agents.

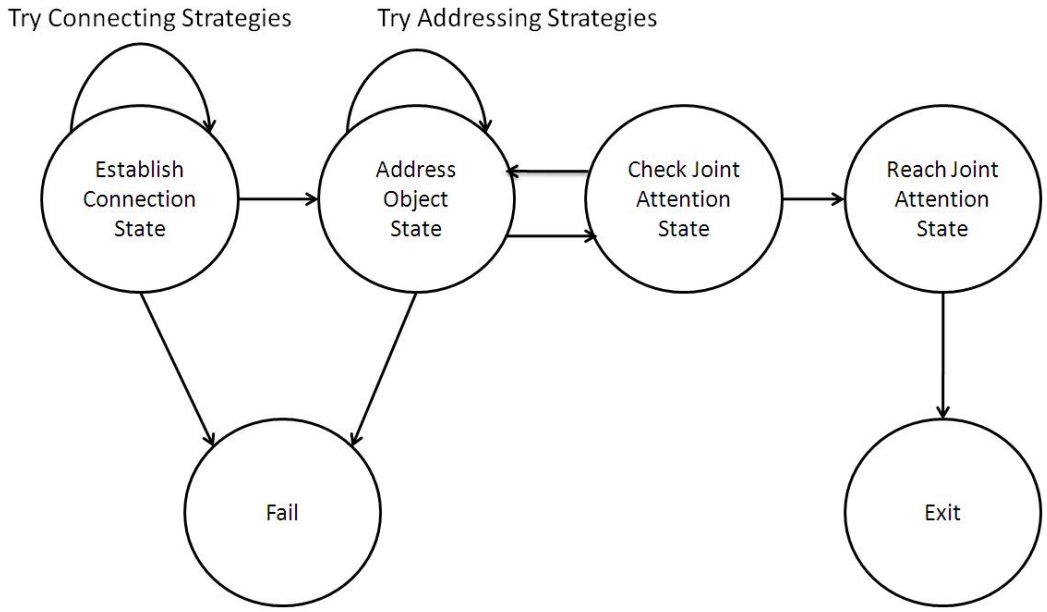

Figure 5: An integrated model of IJA and EJA.

After addressing the focus, the agent checks whether or not joint attention is reached. If not, the agent selects the next available addressing strategy until no strategies are available (i.e., ending in failure to reach joint attention). Otherwise, the agent continues the interaction.

Conceptually, EJA could be viewed as two parts: monitoring and ensuring. In practice, monitoring is the behavior of looking back and forth between the other agent and the referential object, and ensuring is using addressing strategies to make sure joint attention is reached. Moreover, EJA could be categorized as two types based on the time of occurrence: (1) EJA coupled with IJA, which happens right after IJA to ensure its success, and (2) periodical EJA, which happens throughout the interaction to ensure the other agent is still attending to the referential focus.

4.2.4 Implementation

The computational model of joint attention was implemented working with the Creature 6 software system (C6 for short). An earlier version of the creature architecture

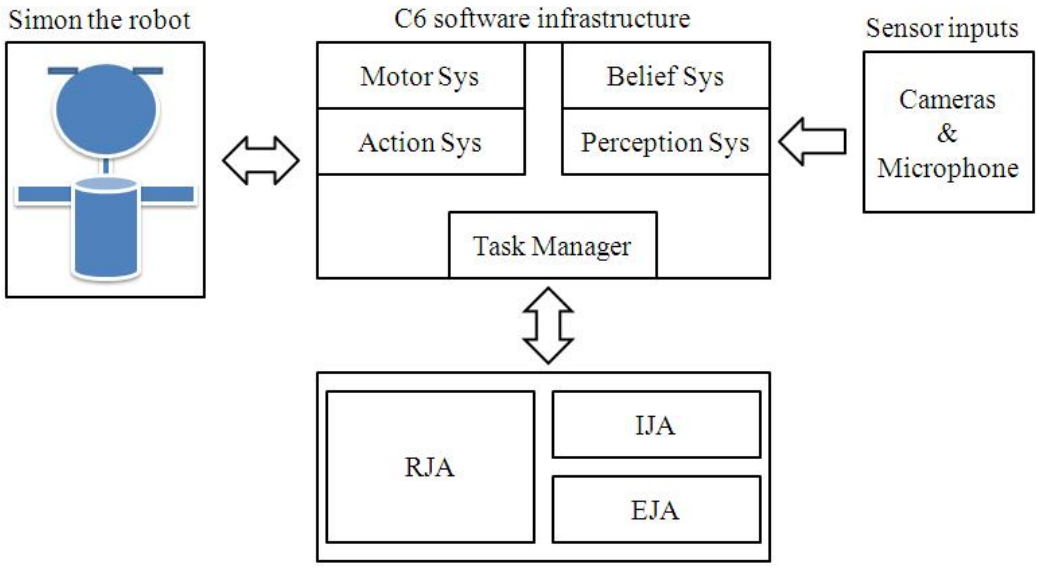

Figure 6: The integrated system of the robot, the software system, and the perception hardware.

(C4) is described in [8]. C6 is a cognitive architecture for interactive agents and follows a sense-think-act construct. It consists of a Perception System, a Belief System, an Action System, and a Motor System. The Perception System takes sensor inputs from the world and constructs this sensor information into beliefs. The Action System arbitrates which action to do according to current beliefs in the Belief System. The Motor System then carries out the selected action by commanding the embodied platform. C6 serves as a software interface to communicate with an embodied platform and perception hardware. Figure 6 shows the integrated system of the robot, the software system, and the perception hardware.

4.3 Embodiment

The benefit of using an embodied platform for evaluation of cognitive capacities has been widely accepted [13, 18]. For the present research, we use a robotic platform to evaluate the proposed computational model of joint attention.

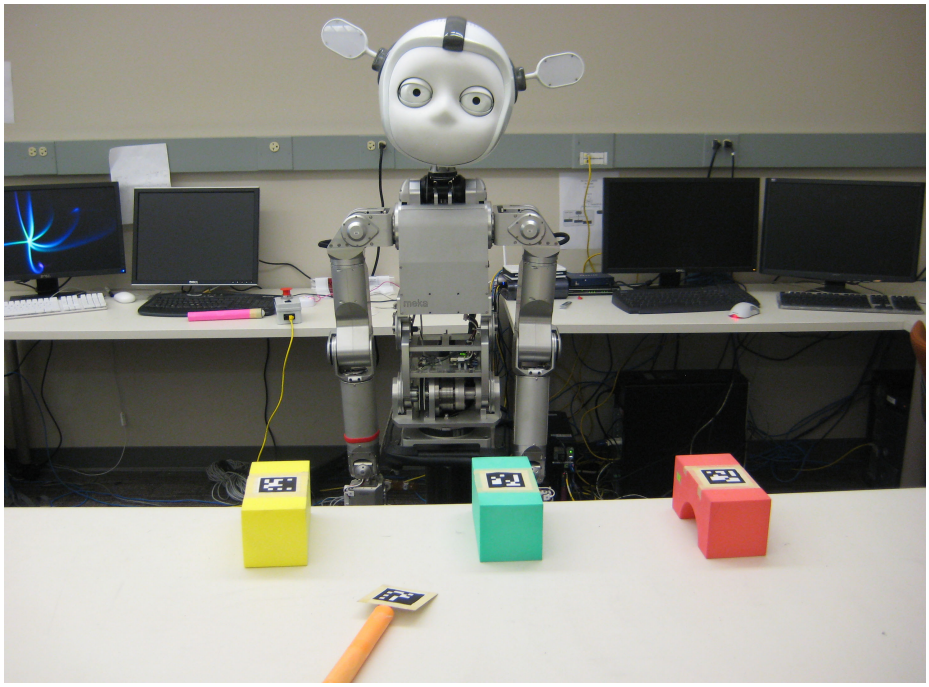

Figure 7: The Simon robot and the RJA experimental setting.

4.3.1 Robotic Platform

The robotic platform for this research is the Simon robot. Simon is an upper-torso humanoid robot with two 7-DOF arms, two 4-DOF hands, and a socially expressive head (Figure 7). Simon has 2-DOF eyes and expressively colorful ears, which can be used as another channel for communication. Simon is capable of turning its head, eyes, and torso as paying attention and showing pointing gestures.

4.3.2 Perception Inputs

We have two external cameras for object recognition, two cameras in Simon's eyes for face tracking, a speech server for speech recognition and synthesis, and a voice volume detector. All the communication among Simon, C6, and perception inputs uses the Intra-Robot Communication Protocol (IRCP) that was developed for communication between software and hardware components.

For object recognition, we used the ARToolKit [1]. Each object that Simon needs to recognize has a predefined fiducial marker associated with it. In addition, we

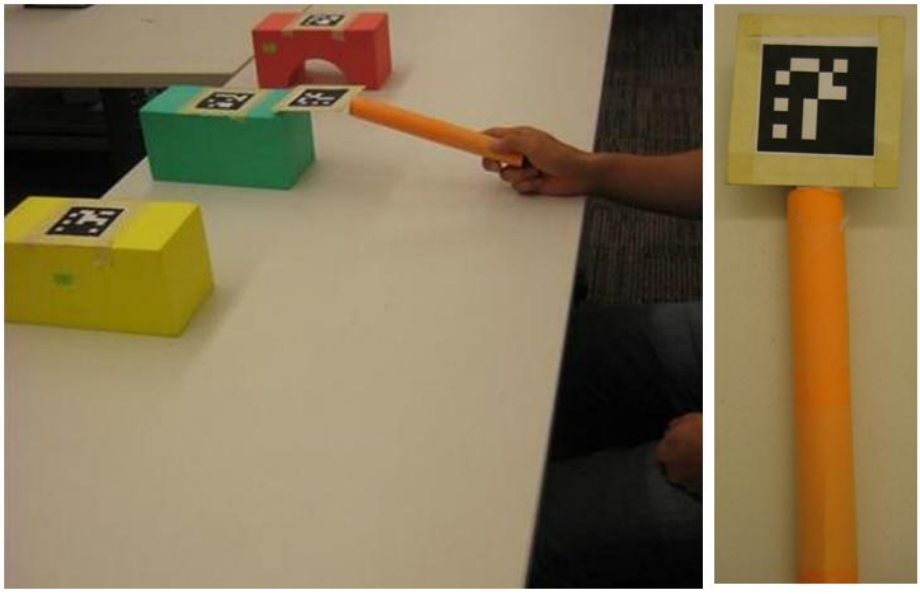

Figure 8: A paper pointer as an index finger for pointing gesture recognition.

also used the ARToolKit for pointing gesture recognition by making a paper pointer (Figure 8(a)) with a marker. Subjects used the paper pointer as their index fingers to point to objects (Figure 8(b)).

Face tracking is accomplished by keeping a detected face at the center of Simon's eye camera view. Therefore, Simon seems always to face to subjects during the interaction. Particularly, we used the face detection utility (Haar Cascade classifier) in OpenCV and applied criteria, such as a reasonable size of a face, to filter out false positive recognitions. When idle, Simon does face tracking to stay engaged with subjects.

For speech, we used the speech technology services in Microsoft robotics developer studio for speech recognition and speech synthesis. A grammar specifying phrases that Simon could understand is pre-defined in XML for speech recognition. In addition to speech sentence detection, PortAudio [2] is used for voice volume detection.

CHAPTER V

EFFECTS OF RESPONDING TO JOINT ATTENTION ON HUMAN-ROBOT INTERACTION

In this chapter, I present an experiment to show the effectiveness of the responding to joint attention (RJA) component and to investigate the effects of RJA in human-robot interaction. This experiment revealed the importance of RJA for robots to effectively interact with humans and a coupling relationship between initiating and ensuing joint attention in interaction.

5.1 Hypotheses

We have four hypotheses (H1-H4) on this experiment investigating effects of RJA in human-robot interaction:

H 1: People have a better mental model of a robot when it responds to joint attention requests.

H 2: People perceive a robot responding to joint attention as more competent.

H 3: People perceive a robot responding to joint attention as more socially interactive.

The first hypothesis tries to see if a robot is more transparent when it responds to joint attention. Transparency helps people's understanding of a robot's attention and internal states that further facilitate better human-robot interaction. The third and forth hypotheses try to reveal that people perceive a robot responding to joint attention in a more positive way, which is important in improving the human-robot relationship.

5.2 Experimental Design

5.2.1 Teaching Task

Subjects were given a task to label objects for Simon. The labeling task is essentially a teaching task where subjects teach Simon two main concepts: color and name. Each main concept contains three sub-concepts: yellow, green, and red for the color concept and banana, watermelon, and apple for the name concept. Each color concept is correspondingly mapped to a name concept (yellow for banana, green for watermelon, and red for apple). The overall task for a subject is to teach Simon the six concepts.

To teach Simon the color concepts, subjects used a pointing gesture and predefined utterance phrases to label objects. To simplify visual recognition, subjects were instructed to use a paper pointer as their index fingers for pointing to the objects that they referred to. The predefined phrase for labeling colors is: *[Simon,] this is a* {*yellow/green/red*} *object.*

In contrast to teaching the color concepts, subjects were not allowed to use the paper pointer while teaching the name concepts. Instead, subjects were instructed to watch Simon and say another predefined phrase (*[Simon,] the* {*yellow/green/red*} *object is* {*a/an*} {*banana/watermelon/apple*}) for labeling.

The reason of the two layers of labels is to make the name concept depend on the color concept. If Simon has not learned the corresponding color concept (e.g., yellow), then it will not be able to know the name concept (e.g., banana). This design is meant to make errors in the interaction more explicit.

The teaching phrase ended when subjects told the experimenter that they felt confident that Simon understood all the concepts. Subjects were told to feel free to re-teach concepts or to request confirmations from Simon if they did not think that Simon understood the concepts. To request a confirmation, they can use an optional phrase:

[Simon,] can you point to the {{*yellow/green/red*} *object*}/{*banana/watermelon/apple*}

After the teaching phrase, subjects were asked to shuffle the objects to a random order and to test Simon each concept once by using the phrase for requesting a confirmation. The whole interaction ended after testing.

The interaction space was organized as shown in Figure 2. Three objects were placed on a table in front of Simon. A subject sat across the table to interact with Simon. Beside Simon was a white board that listed phrases used in the interaction for subjects' reference.

5.2.2 Experimental Conditions

In this experiment, I want to compare a robot with RJA to a robot without RJA and to see how RJA affects performance of an interactive task and people's perception of a robot. I use a between subject design to compare two groups: with-RJA group (experimental group) and without-RJA group (control group).

In the with-RJA group, Simon responds to referential foci by gazing at them. A referential focus could be either a pointed object or a mentioned known object. For example, Simon gazes at the pointed object when a subject teaches a color concept with a pointing gesture. Also, when a subject teaches a name concept (i.e., banana), Simon gazes at the referred object (i.e., the yellow object) if the corresponding color concept has been learned. If a referential concept has not been learned yet, Simon will stay focused on the subject. When a subject requests a concept that has not been learned, Simon communicates uncertainty by gazing over all the objects (from the left to the right and then back to the left). These gaze behaviors are the basic RJA mechanism for Simon to communicate with subjects implicitly.

In the without-RJA group, Simon does not respond to any referential foci. That is Simon does not use its gaze to attend to joint attention that a subject initiated. Therefore, Simon stays focused on subjects throughout the teaching interaction.

However, in both groups, Simon has two basic behaviors. First, Simon always

tracks a subject's face when it does not pay attention to a referential focus. Thus, Simon always tracks a subject's face under the without-RJA condition. Second, Simon's ears blink when it hears any utterance. The intensity of color is modulated by the detected volume. The blinking is not only a way to tell a subject that the speech recognition engine is working but also to make Simon more life-like in terms of social awareness. Note that ear blinking does not mean that Simon understands the concept or what a subject is saying. This was explicitly told to participants.

5.2.3 Procedure

Twenty-four subjects were recruited for this experiment. Four of them were discarded because of speech recognition engine, vision software, or control software crashed during the interaction. All of the valid 20 subjects (19 males and 1 female) were students from the local campus population and were randomly assigned to either the with-RJA or the without-RJA group (10 in each group). A total of seven subjects (four from the with-RJA group and three from the without-RJA group) reported that they did not have any experience related to robotics (including course work, research, or interaction). Most subjects were from computer science or engineering related majors.

First, the experimenter introduced Simon to subjects and mentioned its capabilities (e.g., pointing gesture and blinking ears). The experimenter then introduced a list of phrases that Simon can only understand and presented subjects the task: to teach Simon six concepts, three colors and three names. After the instruction, subjects had a few minutes to get familiar with the phrases and to test voice volume. They started the interaction once they were ready. The interaction process was video recorded for off-line analysis. After the interaction, they were asked to fill out a self-report questionnaire and survey. Finally, the experimenter explained the experiment and answered questions they had. Since the robot interaction is relatively short (average

Table 3: The correlations between the hypotheses and measures for the RJA experiment.

Hypothesis	Measures
H1	M1, M2, M3, M4, Q1, and Q2
H2	Q3, Q4, Q5, and Q6
H3	Q3, Q6 and Q7

less than 15 minutes) subjects were not given compensation for participation.

5.2.4 Measures

To evaluate our hypotheses, we analyzed the interactions from three different angles: quantitative measures of interaction (M1-M4), post-experiment questionnaire (Q1-Q7) and survey (S1-S3), and behavioral observations. Quantitative measures provide objective perspectives on the interaction. Post-experiment questionnaire and survey tell the interaction from a subject's perspective while behavioral observations are from a third-person perspective. Table 3 shows the correlations between the hypotheses and measures.

Quantitative measures of interaction in this experiment are listed as follows.

M 1: Number of errors during the teaching phase: an error is defined as when a human subject either requests a confirmation of a concept that has not been learned or teaches a concept different from the ground true (i.e., labeling the yellow object as an apple)

M 2: Number of interactive steps before recovering errors: this is to measure how many interactive steps between the point where an error occurs to the point where the error is fixed.

M 3: Number of redundant labels during the interaction: a redundant label is when a subject has made the same label attempt before (no matter the concept had been learned correctly or not). Note that a label attempt did

not count as a redundant if it was a repetition due to an utterance not being detected by the speech recognizer.

M 4: Number of confirmations during the teaching phase: this is to measure how many times a subject use the predefined sentence to request confirmations from the robot.

In the post-experiment questionnaire, subjects were asked to rate their perception of the interaction (Q1-Q2) and the robot (Q3-Q7). We used a 7-point rating scale. 1 means most negative; 7 means most positive. For question 6 and 7, 1 means totally disagree and 7 means totally agree.

Q 1: In the interaction, was it clear whether the robot understood what object you referred to?

Q 2: In the interaction, was it clear whether the robot understood the concepts before you requested any confirmation?

Q 3: Was it easy to teach/interact with the robot?

Q 4: Is the robot, in terms of social behaviors, life-like?

Q 5: Is the robot intelligent?

Q 6: Do you agree that the robot will be a good partner in a cooperated task?

Q 7: Do you agree that the robot was engaged in the labeling task?

In the self-report survey, subjects were asked the following questions.

S 1: How do you describe the robot in terms of social behaviors?

S 2: What did you think/do when the robot responded to you?

S 3: What did you think/do when the robot did NOT respond to you?

Table 4: Results of quantitative measures for the RJA experiment. All measures are significant less in the RJA case.

	with-RJA n=10		without-RJA n=10		Significant level	
	Mean	S.D.	Mean	S.D.	t	p<
M1: errors	0.2	0.42	2.9	2.51	4.98	0.001
M2: recovering steps	4	2.83	22.56	35.47	10.27	0.001
M3: redundant labels	2.8	3.74	7.8	10.69	6.00	0.001
M4: confirmations	4.5	1.18	9.8	5.98	6.27	0.001

5.3 Results

5.3.1 Quantitative Results

Table 4 summarizes results of the quantitative measures. Subjects have a better mental model of a robot if it has transparent channels of social communications (e.g., responding to joint attention). For example, a robot uses RJA to convey its understanding of the concept taught by a subject. There was a significant difference between the compared conditions on M1 ($t(18)=4.98$, $p < 0.001$, see Figure 9(a)). On average, 2.9 errors occurred in the without-RJA group while on average only 0.2 errors happened in the with-RJA group. From observation of the interaction, we believe this difference was mainly from subjects in the without-RJA group tending to teach the robot at their own pace, which was usually too fast for the robot to process. Whereas in the with-RJA group, the robot's reactions helped to slow the pace to its desired level. Therefore, more errors were generated during the interaction where the robot did not use RJA.

There was also a significant difference on M2 ($t(18)=10.27$, $p < 0.001$, see Figure 9(b)). It took longer for subjects in the without-RJA group to identify and to correct errors (22.56 interactive steps on average). In contrast, the with-RJA group needed average 4 steps to correct errors. Without RJA on the robot, subjects had hard time to tell if the robot had learned the concept or not. From observation of the

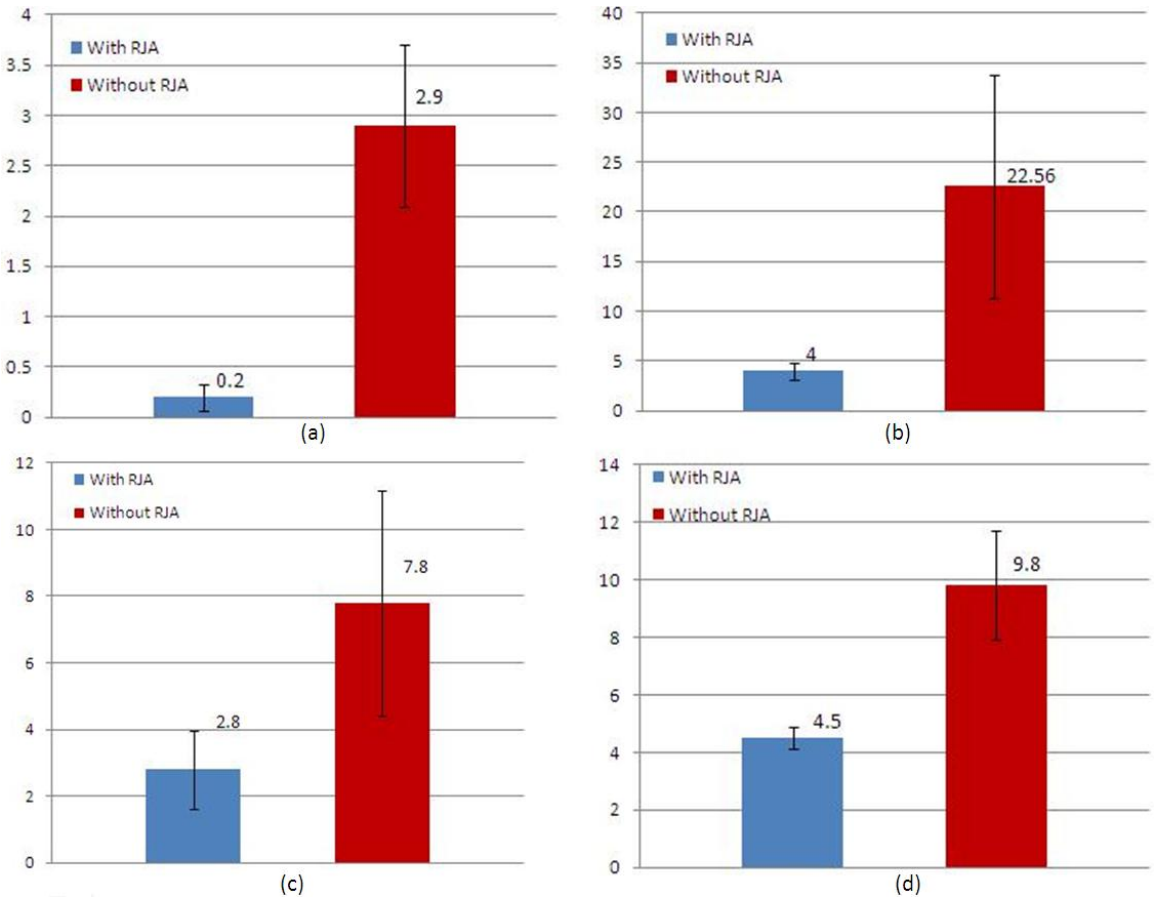

Figure 9: The quantitative results of the interaction of the RJA experiment. (a) is the comparison of number of errors during the teaching phase. (b) is the comparison of number of steps for correcting errors. (c) is the comparison of number of redundant labels during the interaction. (d) is the comparison of number of confirmations during the teaching phase.

interaction, we found that if a subject tended to believe that Simon had learned a concept even without responses from Simon and in fact Simon had not learned the concept, then it took longer for the subject to figure out existence of an error and correct it. This evidence showed that RJA served as a good transparency for subjects to understand the robot.

Similarly, by getting responses from the robot, subjects in the with-RJA group were more confident that the robot had learned the concepts. This was supported by findings on M3 that subjects in the without-RJA group had significant more redundant labels than in the with-RJA group (7.8 times versus 2.8 times per person) ($t(18)=6$, $p < 0.001$, see Figure 9(c)). Without responses from Simon, subjects tended to label multiple times to ensure that Simon had learned the concepts.

Additionally, subjects in the without-RJA group requested more confirmations from Simon until they felt confident that Simon had learned the concepts (average 9.8 times per person) ($t(18)=6.27$, $p < 0.001$, see Figure 9(d)). It is worth to note that if a subject requested a confirmation for each concept it needed 6 times. However, in the with-RJA group, subjects requested less (4.5 times) than this baseline. This showed that subjects in the with-RJA group had a better understanding of the robot's internal states (i.e., had learned the concepts or not) and are more confident in the capacity of the robot.

In summary, the results on M1, M2, M3, and M4 supported the H1 hypothesis that people have a better mental model of a robot that responds to joint attention requests. RJA serves as a transparent communication. With a transparent communication channel between a person and a robot, performance of an interactive human-robot task is better.

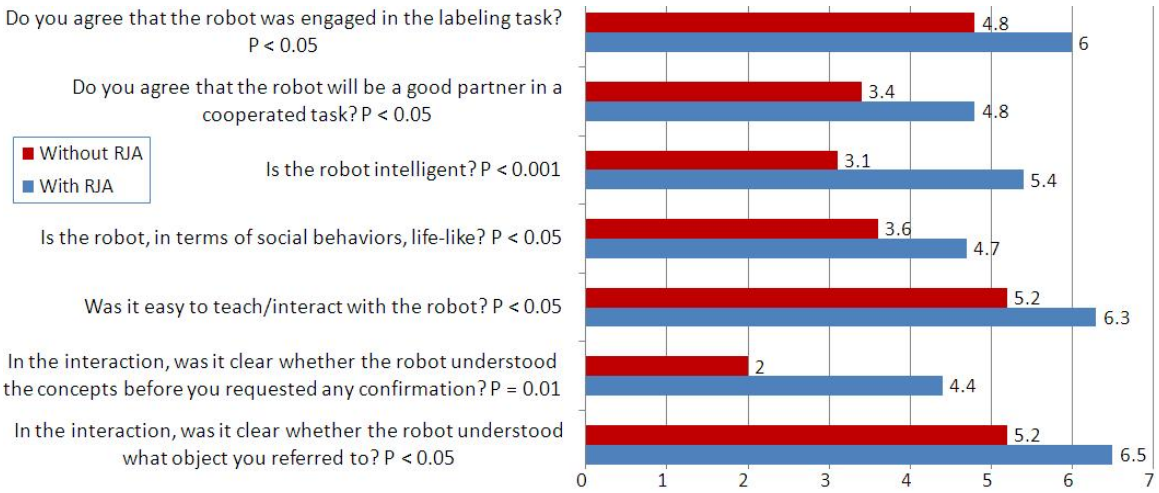

Figure 10: Results of the post-experiment questionnaire for the RJA experiment.

5.3.2 Questionnaire and Survey Results

In addition to the quantitative measures of the interaction, the post-experiment questionnaire shed insight on how subjects perceived of the interaction (Q1 and Q2) and the robot (Q3-Q7). Figure 5.3.2 summarizes the results from analysis of the post-experiment questionnaire. Regarding to perception of the interaction, subjects in the with-RJA group rated that it is much clearer that the robot understood which object they referred to ($t(18)=2.27$, $p < 0.05$) and much clearer that the robot understood the concepts before any confirmations ($t(18)=3.9$, $p = 0.01$). Moreover, they reported it was much easier to interact with the robot ($t(18)=2.26$, $p < 0.05$). Regarding to perception of the robot, subjects in the with-RJA group rated the robot was more life-like in terms of social behaviors ($t(18)=2.5$, $p < 0.05$) and intelligent ($t(18)=4.8$, $p < 0.001$). Additionally, they agreed that the robot will be a good partner in a cooperated task ($t(18)=2.48$, $p < 0.05$) and that the robot was engaged in the interaction ($t(18)=2.34$, $p < 0.05$).

The results from the questionnaire not only support H1, but also uphold hypotheses H2 and H3 that people perceive a robot responding to joint attention in a more positive way.

We examined subjects' responses to the self-report survey (S1-S3) and found those responses were also consistent with the quantitative results and the questionnaire results. Table 5 lists some representative responses. On the one hand, subjects in the with-RJA group were likely to recognize the robot's head or eyes movement as social behaviors and to use them to maintain a mental model of the robot. Moreover, four subjects in the with-RJA group reported that the robot always responded, and two reported that the robot not responding may be due to the poor speech recognition. On the other hand, the without-RJA group reported that they felt awkward and frustrated when the robot was not responding to them. Furthermore, due to lack of responses from the robot, some subjects reported that it was not clear whether the robot did not understand the concept or that they were not speaking clearly.

Table 5: Representative responses from the self-report survey for the RJA experiment.

	How do you describe the robot in terms of social behaviors?	**What did you think/do when the robot responded to you?**	**What did you think/do when the robot did NOT respond to you?**
With RJA	"The robot used head/neck/eye movements to convey focus, and blinking/ear-lights to convey awareness." "It is capable of moving its head direction and convey to the user to which object it is attending. When it is not attending to any object, it keeps looking at me for any instructions, which is natural in this scenario where I am teaching the robot something." "The head movements are good to see that he knows the location of the object I was referring to."	"the eyes were my cue" "Once noticing that its eyes followed my suggestions, I watched its eyes to verify that it was focusing on the object which I was focusing on."	"Troubleshooted the robot like a computer. Performed test cases from simple to complex." "The robot always responded (either by an action when requested to do something, or by showing it had heard me when instructed in something)"
Without RJA	"It has certain very basic motions and behaviors, but socially seemed unaware of its surroundings or focused on something else." "It's very strange; socially, most people don't just stare at you the entire time. I did think it was helpful when it looked at what it was going to point at, though."	"Hard to read the signals. A more obvious indicator would be better." "I was a little bit frustrated when the robot didn't respond sometimes. "	"It was frustrating, because it felt like the robot was ignoring me." "It should give a more clear indication when it does not understand." "I was a little confused whether it didn't respond because it didn't understand what I said or because it didn't know where the referenced object was" "I had a tough time telling why it wasn't doing anything. I didn't know if I wasn't speaking clearly enough, whether or not the robot wasn't listening, or if the robot just didn't understand the concept."

5.3.3 Behavioral Observations

Video analysis on behaviors of subjects also confirmed the hypotheses and gave insights to natural human-robot interaction. Three interesting observations were common across both groups. First, subjects looked back and forth between the referred object and Simon's face to see if Simon understood the concepts. This behavior was actually ensuring joint attention (EJA). Even though not part of this study explicitly, this observation supports overall thesis that EJA is needed in natural human-robot interaction. Second, subjects showed RJA (i.e., followed Simon's pointing gesture) when Simon initiated a joint attention (i.e., pointed to the object of request). This observation revealed that it is natural for humans to attend to or respond to joint attention initiated by the other partner even when that partner is a social robot. Third, subjects tended to give positive or negative responses to the robot. For example, when Simon pointed to the right object, subjects nodded, smiled, laughed, or even said positive words, such as "good." Also subjects frowned or said negative words such as "no" when Simon pointed to a wrong object. These observations may indicate that humans naturally tend to give responses or feedbacks to a learner either explicitly (i.e., utterance) or implicitly (i.e., facial expressions).

Subjects expected responses, especially from the face, from Simon during the interaction. In the first couple of interactions, subjects in the without-RJA group tended to hold a pointing gesture for a relatively long time and waited for responses from Simon. Even after figuring out that Simon did not respond to joint attention, subjects still kept looking at Simon's face. Furthermore, some subjects in the without-RJA group turned to the experimenter and asked if Simon was working properly after their first labeling attempt. These observations were consistent with the self-report survey that subjects felt strange and awkward when Simon did not respond to them. After noticing that Simon lacked responses, subjects used different actions trying to get attention from Simon. Some subjects used additional actions (e.g., tried to move

the objects) while some emphasized the pointing gesture by double-pointing.

Eye gaze served as a good turn-taking mechanism. In the with-RJA group where Simon used eye gaze to respond to joint attention, subjects knew better when their turn to teach a concept was and when the time for Simon to learn a concept was. In particular, subjects waited until Simon looked back from the object to them to give another command. Also subjects tended to hold the paper pointer until Simon looked back at them. In contrast, subjects in the without-RJA group had no clue as to when their turns were. Therefore, subjects tended to teach Simon at their own pace, which is not only an unnatural pace but also usually too fast for Simon to process information correctly. Thus, there were more errors during the interaction and subjects needed more steps to correct errors.

5.4 Summary

Subjects used Simon's eye gaze as a main channel to access Simon's mental model (i.e., whether or not it understood the concepts or where it was focusing on). This claim was proved by the fact that subjects in the with-RJA group requested significant less confirmations and committed significant less errors during the teaching phase. Moreover, some subjects in the with-RJA group explicitly reported that they used Simon's eyes as a cue and to verify if Simon focused on the right object. Subjects have hard time determining whether Simon did not understand the concepts or speech recognition did not recognize correctly if Simon did not show RJA. When Simon did not respond to requests, subjects tended to re-teach the concepts which resulted in redundant labels. This phenomenon is supported by the quantitative findings that the without-RJA group has more redundant labels and needs more steps to correct errors. The results also confirm one prior work on nonverbal study with Leo, where people went too fast, and eye gaze was good for getting people to notice errors early and correct them [10]. Overall, these results supported our hypothesis that people

have a better mental model of a robot when it responds to joint attention requests. Moreover, results from the questionnaire and survey upheld our hypotheses H2 and H3 that people perceived robots responding to joint attention more competent and socially interactive.

CHAPTER VI

THE IMPORTANCE OF ENSURING JOINT ATTENTION IN HUMAN-ROBOT INTERACTION

In this chapter, I present two aspects of the importance of ensuring joint attention (EJA) in human-robot interaction: task performance and naturalness of behavior. Interactive or cooperative tasks can easily break down when one agent is distracted and the other agent does not ensure joint attention. I start with a case study to illustrate the role of EJA in human-robot interaction. I then report an experiment revealing the importance of ensuring joint attention in generating natural interactions and in succeeding in tasks.

6.1 A Case Study: A service robot

Let us consider a service robot in a reception scenario. The main task of the robot is to receive a guest at the reception desk. Instead of dedicated to one task, a service robot capable of multiple tasks is more desirable. Thus, in addition to the main reception task, the robot also has a secondary task of watering plants. Suppose a guest comes to the reception desk and asks if one of her friends, Bob, is here. The robot has the guest wait a moment and turns to deliver a message "There is a guest for you" to Bob. Suppose Bob sits in front of a computer focusing on his work when the robot comes to him. The robot gives a prompt "excuse me, sir" to get Bob's attention (i.e., try to establish a connection). Bob hears and turns to the robot. Unfortunately, Bob accidentally drops of a cup of coffee while turning around. He is distracted and tries to clean up the mess before continuing the interaction with the robot (maybe because the carpet is valuable). Suppose the robot does not ensure

joint attention. In this case, the robot would just deliver the message no matter if Bob is listening or not, and go back to continue its secondary task. Thus, Bob may not actually receive the message that he has a guest waiting. He may need to go to the robot and ask for the message again or he may just forget and keep the guest waiting. Either result shows that the task performance of the robot drops or that the task fails without EJA. If instead, the robot ensures Bob was attending to the joint attention and actually received the message, it will be more effective. In this case, the robot should wait until Bob finished the cleaning.

6.2 Hypotheses

As illustrated in the case study, we hypothesize that ensuring joint attention affects performance in an interactive task. Moreover, psychological findings [24] and our observations on human-robot interaction (see chapter 5) drive us to believe that ensuring joint attention is a natural behavior that humans do. Therefore, we have two hypotheses (H1 and H2) as follows:

H 1: When a robot ensures joint attention it yields better interactive task performance.

H 2: Ensuring joint attention is perceived as a natural behavior in social interaction with a robot.

6.3 Experimental Design

6.3.1 Task and Scenarios

To test our hypotheses we ran a video-based experiment. Subjects were given a task to rank a collection of videos where Simon used varying degrees of ensuring joint attention in three different scenarios. The first scenario (presentation) was Simon as a tour guide robot, giving a presentation to a person. Simon stood beside a poster and faced a person to give a presentation. First, Simon greeted the person and

gave a brief introduction about the presentation. When Simon was about to start the presentation, the person's cell phone rang. The person took the phone call and walked away. Once finished the phone call, the person walked back to re-engage in the presentation. The second scenario (reception) was Simon as a service robot, receiving a guest at a reception desk. This scenario was described in section 6.1. For the purpose of testing effects of EJA on task performance, in both presentation and reception scenarios, the person in the videos was distracted by external events such as getting a cell phone call and dropping a cup of coffee. The third scenario (directions) was Simon as a guide robot, directing a person to the restroom in a building. A person came to Simon and asked where the restroom is. Simon answered with directional speech and a directional gesture.

All scenarios were designed to depict human-robot interaction in real circumstances. We can envision that the presentation scenario will be a common task of a tour guide robot. Consider two applications. First, a robot gives tours in a museum or exhibition. This task is to have the robot move around and give several presentations. Second, a robot promotes a product in a shopping mall. This task is essentially the same as our presentation scenario except that it may involve more interactions (i.e., conversations) between the robot and customers. Similarly, the reception scenario presents a general case to service robotics where a robot has one or two main tasks and several secondary tasks. People have envisioned that service robots in personal houses or in public (e.g., department stores) will help them with routine work. Finally, the directions scenario represents a general task for a guide robot. One of the basic functionalities that a guide robot should have is to direct people to give directions.

Table 6: Behavioral variations in the presentation scenario for the EJA experiment.

Variations	EJA+IJA		Periodical EJA
	Monitoring	Ensuring	
Presentation:V1	√	√	√
Presentation:V2	√	×	√
Presentation:V3	×	√	×
Presentation:V4	×	×	×

6.3.2 Experimental Conditions

Recall that EJA consists of two parts: monitoring and ensuring. In addition, EJA has two different types. One type is EJA coupling with initiating joint attention (IJA), and this type of EJA occurs right after an agent initiates a joint attention. The other type is EJA itself occurring periodically during an interaction. This type of EJA is to maintain joint attention between interacting agents in an interaction.

For the presentation scenario, we studied both parts and types of EJA. Four different behavioral variations are listed in Table 6 (we assumed periodical EJA is the same as monitoring behavior in this scenario). One manipulating variable was whether Simon did EJA after IJA. In videos, Simon exhibited varying degrees of EJA behaviors (i.e., behavioral variations) in response to the distracting event (i.e., a cell phone call). In particular, to ensure joint attention Simon waited until the person finished the phone call and then started the presentation. The other variable that we manipulated was Simon whether or not did EJA periodically during the interaction. We expected people would prefer *Presentation:V1* over the other behavioral variations with respect to task performance and naturalness of behavior.

For the reception scenario, we tested the two parts of EJA. Three different behavioral variations are listed in Table 7. We removed the behavioral variation where ensuring joint attention occurs without monitoring joint attention because this behavior is infeasible in reality. To determine whether or not the other agent is attending

Table 7: Behavioral variations in reception scenario for the EJA experiment.

Variations	EJA+IJA	
	Monitoring	Ensuring
Reception:V1	√	√
Reception:V2	√	×
Reception:V3	×	×

to the joint attention, one agent should do a monitoring behavior first (i.e., look back and forth between the other agent and the referential object). Likewise, in this scenario, Simon exhibited varying degrees of EJA behaviors in response to the distracting event (i.e., dropping a cup of coffee). In *Reception:V1*, Simon monitored Bob's attention and noticed Bob was distracted. It then waited until Bob cleaned up to ensure joint attention was reached. In *Reception:V2*, Simon monitored Bob's attention and noticed Bob was distracted, but it seemed to ignore Bob's situation and just forwarded the message "There is a guest for you." In *Reception:V3*, Simon did not notice Bob's situation and just forwarded the message. We expected people would prefer *Presentation:V1* over the other behavioral variations with respect to task performance.

For the directions scenario, we tested the effect of periodical EJA during an interaction. Two behavioral variations are listed in Table 8. In this scenario, Simon always executed EJA after IJA fully (i.e., monitoring and ensuring). The only variable we manipulated was whether Simon did EJA occasionally or not during the interaction of directing the person to the restroom. In *Directions:V1*, Simon looked back to the user during the interaction to ensure the user was still paying attention to it. In *Directions:V2*, Simon simply gave directions without looking back to the user. We expected *Directions:V1* is more desirable in terms of naturalness of behavior.

Finally, we use a within-subjects design to measure how people perceive effectiveness of communication and naturalness of behaviors of a robot in human-robot

Table 8: Behavioral variations in directions scenario for the EJA experiment.

Variations	EJA+IJA		Periodical EJA
	Monitoring	Ensuring	
Directions:V1	√	√	√
Directions:V2	√	√	×

interaction. To minimize order effects on results, we randomly sorted the videos into three different groups (i.e., three different orders).

6.3.3 Procedure

Fifteen subjects were recruited for this experiment. All 15 subjects (9 males and 6 females) were students from the local campus population and were randomly assigned to one of the three groups (five subjects in each group). Subjects were from computer science, engineering related majors, economics, MBA, and industrial design backgrounds. A total of seven subjects reported that they had not had experience related to robotics (including course work, research, or interaction) before the experiment.

Subjects were first welcomed to participate in the experiment. The experimenter then introduced the task of watching a collection of videos and ranking the videos according to their observations. Subjects were instructed to a website containing the videos and were told not to forward or skip videos at the first time of watching because there were only slight differences between videos. After the video-watching session for each scenario, subjects were asked to fill out a survey regarding to that scenario. Once they finished the survey, they continued to the video-watching session of the next scenario, and so on. Subjects were allowed to watch videos as many times as they wanted at any point in the experiment. Afterwards, the experimenter explained the experiment and answered questions subjects had. Since the experiment is relatively short (about 20 minutes) subjects were not given compensation for participation.

6.4 Results

6.4.1 Quantitative Results

For both the presentation and the reception scenarios, subjects were asked (1) how well the person in the videos can recall or receive the information from the robot and (2) how good the robot was at communicating information. An example of the distribution of people rating for different degrees of EJA and how well the person in the videos received information is shown in Table 9 (Please refer to Appendix A for all the other data). We used the chi-square test for goodness of fit to test subjects' first choices to each question. The null hypothesis was that the distribution of people's votes on varying EJA behavior variations is even with respect to those questions. The null hypothesis suggests that varying EJA behaviors will not affect people's perception of the robot in terms of task performance and naturalness of behavior. We used the chi-square test to test if the real distribution is significantly different from the even distribution. The significant difference tells us that EJA behaviors would actually affect people's perception of the robot. The results indicated that the full EJA behavior (i.e., Presentation:V1 and Reception:V1) is the most desirable behavior with respect to the two questions (both significant level at 0.01). The result supported H1 that a robot ensuring joint attention yields better performance in an interactive task. In our scenarios, better performance came from better communications, which is also true in interactive tasks in general.

For both the presentation and the directions scenarios, subjects were asked how well Simon engaged the person in the videos. The result showed that the full EJA behavior (i.e., Presentation:V1 and Directions:V1) is the most desirable behavior with respect to engaging the other agent (the chi-square test for goodness of fit, significant level at 0.01). In addition, subjects were asked to rank the videos according to how similar the robot's behaviors are to theirs if they were asked to perform the same task. The result revealed that the full EJA behavior is the most similar behavior

Table 9: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the user in videos attained information

Degree of EJA	How well the person received information 1: the best, 4: the worst			
	1	2	3	4
Presentation:V1	12	2	1	0
Presentation:V2	1	3	9	2
Presentation:V3	2	10	3	0
Presentation:V4	0	0	2	13

to theirs (the chi-square test for goodness of fit, significant level at 0.01), suggesting that full EJA is more natural to humans. The results supported H2 that ensuring joint attention is a natural behavior that people do in interaction.

Furthermore, for all three scenarios, subjects were asked to rank the videos according to their preference if they were asked to design behaviors for a robot in similar scenarios. For both the presentation and reception scenarios, 14 out of 15 subjects agreed that the full EJA behavior is the most desirable one, while 13 subjects agreed the same for the directions scenario (the chi-square test for goodness of fit, all significant level at 0.01). This result did not directly support our hypothesis on naturalness of behavior. However, the result that people would like to have EJA behaviors on a robot may imply people perceive EJA behaviors as more affective and natural behaviors.

6.4.2 Descriptive Results

In the survey of each scenario, subjects were asked to comment on the differences they observed and how they liked or disliked the videos. These comments give us insight that it was in fact the ensuring joint attention behaviors that were playing into people's rankings and choices.

For the presentation scenario, all subjects commented that they noticed the difference where Simon made sure the person was paying attention before the presentation

versus not. They noted that "the robot acknowledged the listener's absence and paused" and "waited and made sure that the guest is paying attention before moving on." Moreover, some subjects mentioned that the robot not waiting the person in the video is rude. For example, "its action of just talking while the users weren't present seemed very rude." Similarly, one subject commented that the robot with EJA "shows the respect to the listener." Furthermore, a subject described EJA behavior is what humans will do in similar situation ("the robot waits for the person to come back then continue the speech because that's most likely how a real person would react to such situation.").

Twelve subjects noted the other difference where Simon looked at the user occasionally versus not. Subjects described that the robot "tried to watch the user in the eye", "looked back at the human, possibly trying to recapture attention", and "intermittently engages the guest." In addition, they also agreed that EJA behavior "can improve the communication." Surprisingly, most subjects recognized the role of eye contact in interaction. They reported that they "ranked based on how frequently the robot made eye contact with the human" and that the robot "tried to make eye contacts."

Similar to the presentation scenario, thirteen subjects recognized the difference of whether or not the robot did ensuring joint attention in the reception scenario. Moreover, some subjects remarked the behavior of EJA led to the success of communication. Representative comments are "successfully communicated with Bob" and "Simon did not confirm for Bob's attention, resulting in lost of communication" (Bob is the person in the videos). In addition, one subject thought the robot not ensuring joint attention is out of social norm and explicitly noted that "it (the robot) ignored his situation." and "seems very intentional and rude." Some subjects even perceived the robot as just a machine that reads a script. For example,"spits out the script (ignoring) whatever Bob is doing" and "the robot simply forwarded the information

without paying attention to whether Bob is able to catch up the information."

However, only five subjects explicitly commented about the other difference of monitoring the behavior of the robot ("Simon gestured with his head that a guest was waiting by looking back and forth between Bob and the guest"). I think the reason for dropping numbers from the presentation scenario to the reception scenario is because the setup and the interaction of the reception scenario are more complicated than the presentation scenario. Therefore, subjects might miss the subtle difference in monitoring behavior.

In the directions scenario, thirteen subjects noticed the difference was whether or not the robot turned to the person during interaction. Some subjects applied meanings to the behavior. For instance, they commented that "it was engaging the user more to look back, as if it say *"I know this is a long answer, but please pay attention to me."*" and "Turning to make sure that they understand." Twelve out of 13 subjects commented this behavior in a positive way. For example, "good communication", "engaged with the guest more", and "is mostly how normal people would behave." However, one subject described the behavior as "unnecessary head turns", showing an alternative perspective. Finally, one subject contributed body language (including head movement and gestures) to interaction and communication ("Body language is an important sign in communication and helps people retrieve and remember the information").

6.5 Summary

H1 and H2 were supported by the questionnaire data. People believe that EJA behaviors improve communication which further improves task performance. Additionally, people perceive EJA behaviors as natural behaviors that humans do and would like to design robots to have EJA behaviors for facilitating human-robot interaction.

Moreover, subjects' comments on the videos supported data from the questionnaire. Most subjects recognized the importance of EJA in task performance and communication. Subjects also noted that it is rude for a robot not to ensure joint attention when the other partner is distracted. In addition, subjects agreed that periodical EJA during an interaction is a natural behavior to re-engage the other partner.

CHAPTER VII

FUTURE WORK

There are still several issues that we have not addressed in the current model. First, when and how frequently should a robot need to do ensuring joint attention (EJA) (i.e., monitoring and ensuring) in an interaction? Even though results of the second study have suggested that EJA behavior is natural, we believe that it is natural only when it happens at the time and frequency that meet people's expectation. Second, the model needs to consider dynamics of crowds to handle interactions with a group of people. We believe that interaction with a person is quite different from interaction with a group of people. For example, instead of ensuring everyone in the group is paying attention, a robot may just need to engage most people in the group. In addition, the strategies for getting attention from a group may be different. More studies are needed to explore dynamics of crowds to adapt the model. Third, a robot should be able to learn strategies through interactions with humans and use strategies adaptively according to situations and the person it is interacting with. Humans have different ways to interact with different people in different situations. To be in a human environment, robots need the ability to adapt themselves from a person to another person and from a situation to another situation.

A couple of milestones in infant development could be included in our model. Around 12 months, infants are able to turn to sounds coming from behind. This developmental milestone could be achieved by using sound localization technique. Obviously, this skill of paying attention to sound facilitates joint attention in human-robot interaction. For example, making sound is a good way to draw attention from the other agent. Another important developmental milestone is the ability to break

gaze and take over control in interaction (i.e., turn-taking). Turn-taking is a crucial key to social learning, such as imitative learning and active learning, in contingent interactions. Robots with turn-taking skill know when to imitate behaviors and when is a right time to ask questions to a human teacher. To achieve turn-taking skill, a robot needs a contingency detector to infer when it can lead the interaction.

Moreover, as suggested in [19], an agent responding to joint attention should also need to occasionally look back to the other agent to check if the initiating agent is still focusing on the referential object. This implies that there may be also an *EJA* mechanism residing in a responding agent. We hypothesize that EJA coupling with responding to joint attention (RJA) has two purposes. One purpose is to ensure self is continuously attending to the right focus that the initiating agent is focusing on. This behavior of looking back to the initiating agent may be a response to the periodical EJA of the initiating agent. The other purpose is to continue the interaction. For instance, the looking-back behavior of a responding agent signals the other agent that "I know what you are talking about, and please continue." However, a human-human interaction study is needed to verify the existence and functionalities of EJA coupling with RJA.

Furthermore, Kaplan and Hafner noted several challenges in realization of joint attention [18]. They suggested four prerequisites of joint attention: attention detection, attention manipulation, social coordination, and intentional stance. Attention detection and attention manipulation match to RJA and IJA respectively. Social coordination in terms of turn-taking is implicitly realized by RJA (See Chapter 5). However, intentional stance, including the ability to differentiate goals and means, and to apply intentions to others, is not covered in our model. To have intentional stance, we believe that agents would need to be able to empathize others (i.e., simulate the other agent's mind and track the other agent's intentions).

Finally, we hope to include human gaze as input to our model because it has been

known that eye gaze is one of the most important social cues that humans use. We had tried eye-tracking technique to have gaze input to predict a person's attention. Unfortunately, the technique is too limited and unreliable in our experimental settings. For example, a person has to limit her position and motions in front of a robot during interaction so that her gaze can be tracked. This restriction limits interactive scenarios. However, instead of using eye gaze, we can try techniques of head tracking to estimate a person's attentional focus. There have been a lot of efforts to the research of accurately tracking humans' eye gaze and head orientation in real-time. As soon as those techniques get mature for robotics applications, they will benefit much on robotics research, especially on human-robot interaction.

CHAPTER VIII

CONCLUSION

Joint attention, a crucial component in interaction and an important milestone in human development, has drawn a lot of attention from the robotics community recently. Robotics researchers study and implement joint attention for robots with the belief that robots with the joint attention ability can (1) interact with humans in a way that humans and humans interact and (2) better learn from humans through interactions. Not only robotics researchers but also researchers from psychology, cognitive science, and neuroscience are interested in implementing models of joint attention on robots because they believe that embodiment provides different perspectives to understanding joint attention.

Most previous work on realization of joint attention in the robotics community focused on responding to joint attention (RJA) and/or initiating joint attention (IJA) only. RJA is the ability to follow another's direction of gaze and gestures in order to attain common experience. IJA is the ability to manipulate another's attention to a focus of interest in order to share experience. However, to the best of our knowledge, there is no work explicitly addressing the ability to ensure that joint attention is achieved by interacting agents. We refer this ability as ensuring joint attention (EJA) and recognize its importance in human-robot interaction.

The contribution of this work is threefold. First of all, we proposed a computational model of joint attention consisting of three parts: RJA, IJA, and EJA. This decomposition is supported by psychological findings and matches the developmental timeline of infancy. Infants start with the skill of following a caregiver's gaze, and then they exhibit imperative and declarative pointing gesture to get the caregiver's

attention. Importantly, the initiating actions often come with an ensuring behavior that is to look back and forth between the caregiver and the referred object. In our model, RJA and IJA run exclusively, whereas EJA is an always-on process interacting with IJA to ensure that the other agent is attending to the right focus. We do not claim that the design of the proposed model fully reflects joint attention in humans. But, we wish to highlight the relationship between RJA, IJA, and EJA.

Secondly, we conducted an experiment to explore the effects of responding to joint attention on human-robot interaction and found that robots responding to joint attention are more transparent to humans. RJA provides transparency of a robot's internal states. Therefore, people have a better idea of what a robot's current state is and have a better mental model of a robot. The transparency leads to better performance of a human-robot interactive task. In addition, people perceive robots responding to joint attention are more competent and socially interactive. We believe that these positive perceptions of robots will improve human-robot relationship.

Thirdly, we conducted another experiment to study the importance of ensuring joint attention in human-robot interaction. This experiment is to draw attention to the existence and importance of EJA in robotics applications. The results showed that EJA behaviors can yield better performance in a human-robot interactive task. In addition, people perceive EJA behaviors as natural behaviors that humans do and would like to design robots to have EJA behaviors for facilitating human-robot interaction.

APPENDIX A

RANKING DATA IN THE ENSURING-JOINT-ATTENTION EXPERIMENT

Table 10: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the robot in videos communicated information

		Goodness of communication of the robot 1: the best, 4: the worst			
		1	2	3	4
Degree of EJA	Presentation:V1	13	2	0	0
	Presentation:V2	1	2	9	3
	Presentation:V3	1	10	4	0
	Presentation:V4	0	1	2	12

Table 11: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the robot engaged the user in the video

		Goodness of engagement with the user 1: the best, 4: the worst			
		1	2	3	4
Degree of EJA	Presentation:V1	14	1	0	0
	Presentation:V2	0	7	6	2
	Presentation:V3	1	5	8	1
	Presentation:V4	0	2	1	12

Table 12: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to what the robot's actions are most similar to a subject's behaviors

		Similarity of behaviors (1: the most similar, 4: the least similar)			
		1	2	3	4
Degree of EJA	Presentation:V1	13	2	0	0
	Presentation:V2	1	3	9	2
	Presentation:V3	1	10	4	0
	Presentation:V4	0	0	2	13

Table 13: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to what behaviors a subject would like a robot to have

		Desirability of behaviors (1: the most desirable, 4: the least desirable)			
		1	2	3	4
Degree of EJA	Presentation:V1	14	1	0	0
	Presentation:V2	0	4	8	3
	Presentation:V3	1	10	4	0
	Presentation:V4	0	0	3	12

Table 14: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the user in videos attained information

		How well the person received information (1: the best, 3: the worst)		
		1	2	3
Degree of EJA	Reception:V1	15	0	0
	Reception:V2	0	11	4
	Reception:V3	0	4	11

Table 15: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the robot in videos communicated information

		Goodness of communication of the robot (1: the best, 3: the worst)		
		1	2	3
Degree of EJA	Reception:V1	15	0	0
	Reception:V2	0	12	3
	Reception:V3	0	3	12

Table 16: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to what behaviors a subject would like a robot to have

		Desirability of behaviors (1: the most desirable, 3: the least desirable)		
		1	2	3
Degree of EJA	Reception:V1	14	1	0
	Reception:V2	1	11	3
	Reception:V3	0	3	12

Table 17: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to how well the robot engaged the user in the video

		Goodness of engagement with the user (1: better, 2: worser)	
		1	2
Degree of EJA	Directions:V1	13	2
	Directions:V2	2	13

Table 18: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to what the robot's actions are most similar to a subject's behaviors

		Similarity of behaviors (1: more similar, 2: less similar)	
		1	2
Degree of EJA	Directions:V1	13	2
	Directions:V2	2	13

Table 19: Contingency table of frequencies of subjects' preference on interactive behaviors regarding to what behaviors a subject would like a robot to have

		Desirability of behaviors 1: more desirable, 2: less less desirable	
		1	2
Degree of EJA	Directions:V1	14	1
	Directions:V2	1	14

REFERENCES

[1] "ARToolKit." http://www.hitl.washington.edu/artoolkit/ (June/2010).

[2] "PortAudio." http://www.portaudio.com/ (June/2010).

[3] American Psychiatric Association, *Diagnostic criteria for 299.00 Autistic Disorder*, 2000.

[4] Bakeman, R. and Adamson, L. B., "Coordinating attention to people and objects in mother-infant and peer-infant interaction," *Child Development*, vol. 55, no. 4, pp. 1278–1289, 1984.

[5] Baldwin, D. A., "Early referential understanding: Infants' ability to recognize referential acts for what they are," *Developmental Psychology*, vol. 29, no. 5, pp. 832–843, 1993.

[6] Baron-Cohen, S., "Precursors to a theory of mind: Understanding attention in others," in *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (Whiten, A., ed.), pp. 233–251, Cambridge, Massachusetts: Basil Blackwell, 1991.

[7] Baron-Cohen, S., *Mindblindness: An essay on autism and theory of mind*. Cambridge, Massachusetts: The MIT Press, 1997.

[8] Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M. P., and Tomlinson, B., "Integrated learning for interactive synthetic characters," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 417–426, 2002.

[9] Brauer, J., Call, J., and Tomasello, M., "All great ape species follow gaze to distant locations and around barriers," *Journal of Comparative Psychology*, vol. 119, no. 2, pp. 145–154, 2005.

[10] Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., and Berlin, M., "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS*, pp. 383–388, 2005.

[11] Butterworth, G. E. and Jarrett, N. L. M., "What minds have in common is space: spatial mechanisms serving joint visual attention in infancy," *British Journal of Developmental Psychology*, vol. 9, no. 1, pp. 55–72, 1991.

[12] Carlson, E. and Triesch, J., "A computational model of the emergence of gaze following," in *In Connectionist models of cognition and perception II* (Bowman, H. and Labiouse, C., eds.), pp. 105–114, World Scientific, 2003.

[13] Deák, G. O., Fasel, I., and Movellan, J., "The emergence of shared attention: Using robots to test developmental theories," in *In Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, pp. 95–104, 2001.

[14] Hobson, R., Patrick, M., Crandell, L., Garcia Perez, R., and Lee, A., "Maternal sensitivity and infant triadic communication," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 45, no. 3, pp. 470–480, 2004.

[15] Hoffman, M. W., Grimes, D. B., Shon, A. P., and Rao, R. P. N., "2006 special issue: A probabilistic model of gaze imitation and shared attention," *Neural Netw.*, vol. 19, no. 3, pp. 299–310, 2006.

[16] Imai, M., Ono, T., and Ishiguro, H., "Physical relation and expression: Joint attention for human-robot interaction," *IEEE Transaction on Industrial Electronics*, vol. 50, no. 4, pp. 636–643, 2003.

[17] Johnson, C. P., Myers, S. M., and the Council on Children With Disabilities, "Identification and evaluation of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1183–1215, 2007.

[18] Kaplan, F. and Hafner, V. V., "The challenges of joint attention," *Interaction Study*, vol. 7, no. 2, pp. 135–169, 2006.

[19] Kozima, H. and Yano, H., "A robot that learns to communicate with human caregivers," in *In Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, 2001.

[20] Langton, S. R. H., Watt, R. J., and Bruce, V., "Do the eyes have it? Cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50–59, 2000.

[21] Levinson, S. C., *Pragmatics*. Cambridge University Press, 1983.

[22] Marin-Urias, L., Sisbot, E., Pandey, A.K. abd Tadakuma, R., and Alami, R., "Towards shared attention through geometric reasoning for human robot interaction," in *9th IEEE-RAS International Conference on Humanoid Robots*, pp. 331–336, 2009.

[23] Mundy, P., Block, J., Delgado, C., Pomares, Y., Hecke, A. V. V., and Parlade, M. V., "Individual differences and the development of joint attention in infancy," *Child Development*, vol. 78, no. 3, pp. 938–954, 2007.

[24] Mundy, P. and Newell, L., "Attention, Joint Attention, and Social Cognition," *Current Directions Psychological Science*, vol. 16, no. 5, pp. 269–274, 2007.

[25] Nagai, Y., Hosoda, K., Morita, A., and Asada, M., "A constructive model for the development of joint attention," *Connection Science*, vol. 15, pp. 211–229, 2003.

[26] Paulev, P.-E., *Medical Physiology And Pathophysiology*. Copenhagen Medical Publishers, 2000.

[27] Pitman, C. A. and Shumaker, R. W., "Does early care affect joint attention in great apes (pan troglodytes, pan paniscus, pongo abelii, pongo pygmaeus, gorilla gorilla)?," *Journal of Comparative Psychology*, vol. 123, no. 3, pp. 334–341, 2009.

[28] Rich, C., Ponsleur, B., Holroyd, A., and Sidner, C. L., "Recognizing engagement in human-robot interaction," in *HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, (New York, NY, USA), pp. 375–382, ACM, 2010.

[29] Scassellati, B., "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, pp. 176–195, Springer Berlin, 1999.

[30] Slaughter, V. and McConnell, D., "Emergence of joint attention: relationships between gaze following, social referencing, imitation, and naming in infancy," *The journal of genetic psychology*, vol. 164, no. 1, pp. 54–71, 2003.

[31] Sokolov, E., "Higher nervous functions: Orienting reflex," *Annual Review of Physiology*, vol. 25, pp. 545–580, 1963.

[32] Striano, T., Reid, V. M., and Hoehl, S., "Neural mechanisms of joint attention in infancy," *The European journal of neuroscience*, vol. 23, no. 10, pp. 2819–23, 2006.

[33] Tager-Flusberg, H. and Caronna, E., "Language disorders: Autism and other pervasive developmental disorders," *Pediatric Clinics of North America*, vol. 54, pp. 469–481, 2007.

[34] Thomaz, A. L., Berlin, M., and Breazeal, C., "An embodied computational model of social referencing," in *In IEEE International Workshop on Human Robot Interaction*, 2005.

[35] Tomasello, M., "Joint attention as social cognition," in *Joint attention: its origins and role in development* (Moore, C. and Dunham, P. J., eds.), pp. 103–130, Hillsdale, NJ, England: Lawrence Erlbaum Associates, 1995.

[36] Tomasello, M., Call, J., and Hare, B., "Five primate species follow the visual gaze of conspecifics," *Animal Behaviour*, vol. 55, no. 4, pp. 1063–1069, 1998.

[37] Volkmar, F., Chawarska, K., and Klin, A., "Autism in infancy and early childhood," *Annual Review of Psychology*, vol. 56, pp. 315–336, 2005.

[38] Wellman, H. M. and Bartsch, K., "Young childrens reasoning about beliefs," *Cognition*, vol. 30, no. 3, pp. 239–277, 1988.

[39] Williams, J. H., Waiter, G. D., Perra, O., Perrett, D. I., and Whiten, A., "An fMRI study of joint attention experience," *NeuroImage*, vol. 25, no. 1, pp. 133–140, 2005.